Protein design and variant prediction using autoregressive generative models

Year: 2021 | Citations: 335 | Authors: Jung-Eun Shin, Adam J. Riesselman, Aaron W. Kollasch, Conor McMahon, Elana Simon

---

Abstract

The ability to design functional sequences and predict effects of variation is central to protein engineering and biotherapeutics. State-of-art computational methods rely on models that leverage evolutionary information but are inadequate for important applications where multiple sequence alignments are not robust. Such applications include the prediction of variant effects of indels, disordered proteins, and the design of proteins such as antibodies due to the highly variable complementarity determining regions. We introduce a deep generative model adapted from natural language processing for prediction and design of diverse functional sequences without the need for alignments. The model performs state-of-art prediction of missense and indel effects and we successfully design and test a diverse 105-nanobody library that shows better expression than a 1000-fold larger synthetic library. Our results demonstrate the power of the alignment-free autoregressive model in generalizing to regions of sequence space traditionally considered beyond the reach of prediction and design. The ability to design functional sequences is central to protein engineering and biotherapeutics. Here the authors introduce a deep generative alignment-free model for sequence design applied to highly variable regions and design and test a diverse nanobody library with improved properties for selection experiments.